

# Asymptotic frequency of shapes in supercritical branching trees

*Giacomo Plazzotta\* & Caroline Colijn†*

## Abstract

The shapes of branching trees have been linked to disease transmission patterns. In this paper we use the general Crump-Mode-Jagers branching process to model an outbreak of an infectious disease under mild assumptions. Introducing a new class of characteristic functions, we are able to derive a formula for the limit of the frequency of the occurrences of a given shape in a general tree. The computational challenges concerning the evaluation of this formula are in part overcome using the Jumping Chronological Contour Process. We apply the formula to derive the limit of the frequency of cherries, pitchforks and double cherries in the constant rate birth-death model, and the frequency of cherries under a non-constant death rate.

**Keywords:** branching processes, shape frequency, basic reproduction number

## 1 Introduction

Branching processes are widely studied and used to model many biological growth phenomena. Although their first and most direct application has been in the study of evolution and extinction [17], they have also been employed to model infectious disease epidemics [15, 28]. In this context each branching event represents an infection at which a new infectious individual enters the model. The individual's history corresponds to a path in the tree, beginning at the time of infection and ending at a tip which represents the individual's death or recovery.

In recent years, improvements in sequencing technologies have made it possible to detect micro-evolutionary events in pathogens. Pathogen sequence data can be used to infer branching trees which in turn can inform our understanding of the disease's transmission dynamics [8, 18, 29]. The inference of trees from sequence data becomes more challenging as more and more isolates (tips in the tree) are sequenced, presenting significant challenges for the field. Inference relies on tree likelihoods, which are derived from branching processes [9, 27, 20]. The shapes of phylogenetic trees have also been linked to disease transmission patterns [11, 6, 25].

The shape of a tree can be defined, informally, as the tree without considering the associated branch lengths [13]. The shape distribution for the Yule model was studied originally in order to estimate the branching points of a branching diffusion process [10]. Similarly, numerous studies can be found on tree likelihoods and inference, though most exploit the timing of branching events in trees rather than focusing on tree shapes. Moreover, the Yule tree is the most well-studied model, for which there are results describing internal structure and shape distribution [4, 14, 24, 3]. Hence the shape distribution, at least for the homogeneous (Yule) model, can be considered as resolved. For non-homogeneous models however, few results are available in the literature.

The frequency of a shape in a tree is the ratio between the number of occurrences of that particular shape and the number of tips in the tree. Some results regarding shape frequency are available for simple tree models [22, 26, 5]. In more general settings, the limit of the frequency of a shape will

---

\*Imperial College London. Corresponding author: giacomo.plazzotta11@imperial.ac.uk

†Imperial College London

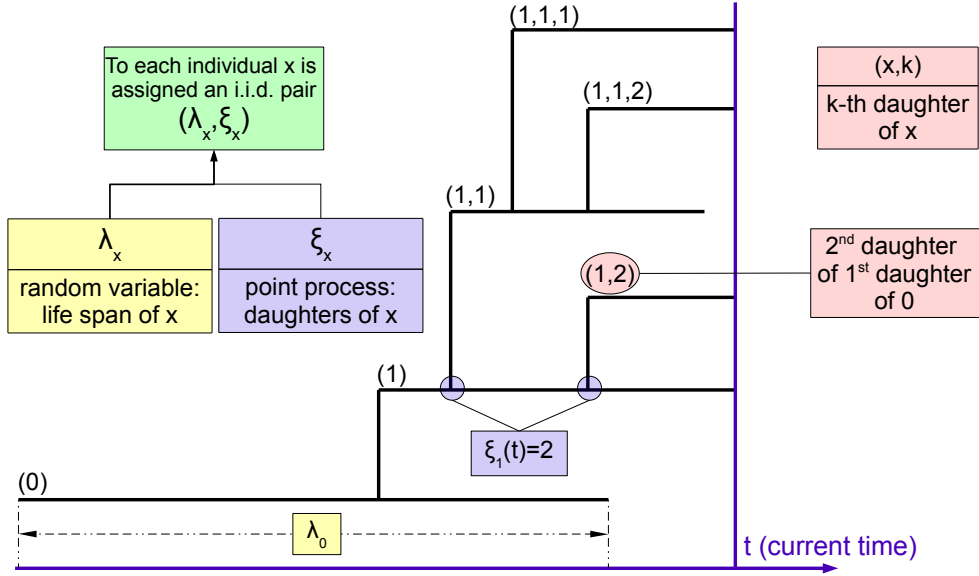


Fig. 1: Explanatory figure that shows the choice of notation in a simple tree generated from a CMJ process

depend on the process defining the tree. Therefore, shape frequencies could provide a tool to estimate the governing parameters of a branching process, using trees derived from empirical data. Given the intractable number of tree shapes with  $n$  tips [14]  $((2n-3) \cdot (2n-1) \dots 3 \cdot 1)$ , computational approaches to finding the frequencies of small shape patterns in large trees can have limited success. This motivates an analytical derivation of shape frequencies.

In this work we use the Crump-Mode-Jagers branching process (CMJ process) [17] to obtain the asymptotic frequencies of tree shapes. The CMJ process is a general model with very mild assumptions; it includes as special cases the Yule process and the homogeneous (constant-rate birth-death) process. We focus on supercritical trees, whose Malthusian parameter is  $\geq 1$ , because they have a positive probability of never reaching extinction. Using novel characteristic functions and previous convergence properties of the CMJ [23], we derive a general formula for the asymptotic frequency of potentially any shape in any tree. The evaluation of the formula presents computational challenges, which we overcome in part by applying the Jumping Chronological Contour Process [19, 21]. We evaluate the expression for the asymptotic frequency of some simple shapes in the homogeneous tree and a non-homogeneous model.

## 2 Background

The theory of general Crump-Mode-Jagers (CMJ) branching processes provides an ideal framework for the study of sub-shapes because they can be counted with characteristic functions. We begin with some definitions and the results we have used from the literature on CMJ processes. More details can be found in [17, 2].

### Notation and definition

Following Jagers' setting [17], each individual of the process is assigned a sequence  $x$  in the space  $I$  of all the possible sequences of non-negative integers. The sequence  $x$  is chosen uniquely in the following way: if the individual with sequence  $x$  is the  $k$ -th daughter of the individual with sequence  $y$ , then  $x = (y, k)$ . Therefore, setting the ancestor's sequence as the singleton 0, each individual's unique sequence keeps track of its predecessors. Since every sequence starts with a 0, for simplicity we will

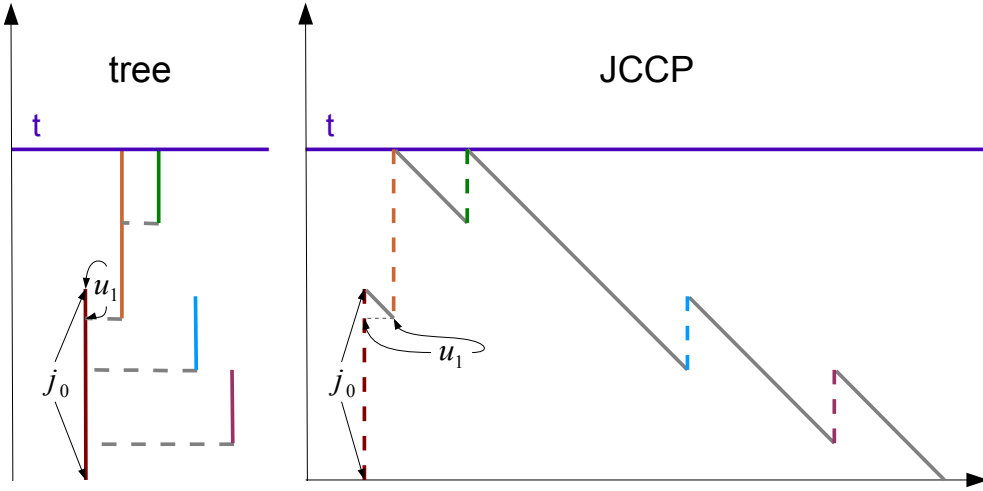


Fig. 2: An example of a Jumping Chronological Contour Process and its relative tree.

omit this leading 0 from the notation.

Each individual  $x$  is also assigned a random variable  $\lambda_x$ , the life length of  $x$ , and a point process  $\xi_x$ , the reproduction of  $x$ . Further, it is assumed that the pairs  $(\lambda_x, \xi_x)$  are i.i.d. but  $\lambda_x$  and  $\xi_x$  may depend on each other. For the scope of this paper we need to make two restrictions on the point process  $\xi$ . First, to have a binary tree, we require that  $\xi$  has no multiple points. Second, let  $\mu(t) = E[\xi(t)]$ : for to have supercriticality of the branching process we require  $\lim_{t \rightarrow \infty} \mu(t) > 1$ .

We refer to the constant-rate birth-death process as the homogeneous process. It is a special case of the CMJ process: births happen at a constant rate  $\beta$  and deaths at a constant rate  $\delta$ . In the CMJ setting, this corresponds to the point process  $\xi$  being a Poisson process with intensity  $\beta$  over the life span  $\lambda$ , which is exponentially distributed with parameter  $\delta$ .

To complete the definition of the branching process, consider the function  $z_x(t)$  that indicates whether individual  $x$  is alive at time  $t$ :

$$z_x(t) = \begin{cases} 1 & \text{if } x \text{ is alive at time } t \\ 0 & \text{otherwise} \end{cases}.$$

Then the CMJ process  $\{Z(t); t \in \mathbb{R}_0^+\}$  is defined as follows:

$$Z(t) = \sum_{x \in I} z_x(t).$$

### Convergence of CMJ processes

Because of the self-similar structure of branching processes, renewal theory is often used to analyze them; we refer to [7, 16, 17] for detailed applications and proofs. Because we want to analyze the asymptotic behaviour of cherries in supercritical branching processes, it is important to state that such processes converge [16], in the sense that:

$$E[e^{-Mt} Z(t)] \rightarrow \frac{\int_0^\infty e^{-Mt} E[z_0(t)] dt}{\int_0^\infty t e^{-Mt} \mu(dt)},$$

where  $\mu(t) = E[\xi(t)]$  and  $M$  is the Malthusian parameter of the process, i.e. the positive real number that satisfies  $\int_0^\infty e^{-Mt} \mu(dt) = 1$ .

### The Jumping Chronological Contour Process

Contour processes can be interpreted as “distance to the root” processes [12], given a proper distance. We use the Jumping Chronological Contour Process (JCCP) defined in [19, 21]. To avoid unnecessary

heavy notation we give an informal description here. Consider a ball that visits every point on the tree starting from the death of the ancestor. The ball proceeds with speed -1 along the lifetime of the individual it is visiting, and when it encounters a birth event (of a descendant of that individual), it jumps to the point representing the death of the newborn. When the ball reaches the birth time of the individual it is visiting, it continues up the tree to visit the mother of that individual (and her descendants, and so on). The process is stopped when the ball reaches the point where the ancestor is born. Figure 2 shows the JCCP and its corresponding tree. The JCCP can be very useful because of the independence of its defining components (jumps and declines) for all trees with constant birth rates.

Construction of the JCCP can be achieved independently from a previously defined tree through simulation of the jumps  $j_i$  and the drops  $u_i$  (see Figure 2). When constructing the JCCP two rules need to be applied: reflection and killing upon hitting 0. The first occurs when a jump overshoots the current time  $t$  and it is reflected or “sent back to” the current time. The second is used to set an end for the process whenever it hits the  $x$ -axis. With these two properties, the JCCP has the same law as the tree [20].

### 3 Characteristic functions for shapes

The shape  $\mathcal{S}$  of a tree or a subtree can be defined as the tree or the subtree without the associated branch lengths [13]. Common small (low number of tips) shapes are the cherry, the three-tip shape known as a pitchfork, the four-tip symmetric shape called a “double cherry”, and so on. If the shape  $\mathcal{S}$  occurs inside a tree, it is clear that there is a unique individual which is both a tip and an ancestor of  $\mathcal{S}$ , for each occurrence of  $\mathcal{S}$  in the tree. If  $x$  is the ancestor of shape  $\mathcal{S}$  we say that  $x$  *mothers*  $\mathcal{S}$ .

**Definition 3.1:** The characteristic of the shape  $\mathcal{S}$  is

$$\phi_x^{\mathcal{S}}(t) = \begin{cases} 1 & \text{if } x \text{ mothers and is a tip of } \mathcal{S} \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

The total number of occurrences of shape  $\mathcal{S}$  in the tree at time  $t$  is  $Z^{\mathcal{S}}(t) = \sum_{x \in I} \phi_x^{\mathcal{S}}(t)$ . For simplicity, if the suffix  $x$  is not specified for the characteristic, we imply the ancestor. Note that the characteristic may not be independent between individuals, but we require that it only depends on  $x$ ’s life and on its daughter process. Henceforth we shall assume the following conditions:

$$\sum_{k=0}^{\infty} \sup_{k \leq t \leq t+1} (e^{-Mt} E[\phi(t)]) < \infty, \quad (2)$$

$$E \left[ \sup_{s \leq t} \phi(s) \right] < \infty, \text{ for all } t < \infty. \quad (3)$$

Statements (2) and (3) were used in [23] to prove a number of convergence properties of supercritical trees. In general the number of occurrences of a particular configuration (shape) does not converge in a supercritical tree. However, given two shapes  $\mathcal{S}^1$  and  $\mathcal{S}^2$  with characteristics  $\phi^1$  and  $\phi^2$  satisfying equations (2) and (3), the ratio of the numbers of these two shapes does converge. From [23] we have:

$$\frac{Z^{\mathcal{S}^1}(t)}{Z^{\mathcal{S}^2}(t)} \rightarrow \frac{\int_0^{\infty} e^{-Mt} E[\phi^1(t)] dt}{\int_0^{\infty} e^{-Mt} E[\phi^2(t)] dt}, \text{ in probability as } t \rightarrow \infty. \quad (4)$$

In equation (4) the characteristics  $\phi^1$  and  $\phi^2$  are associated with the ancestor, i.e.  $\phi^1 := \phi_0^1$  and  $\phi^2 := \phi_0^2$ . This is to ensure that the time variable in the characteristic and in the integral are the same, without delays.

The fact that the convergence expressed in equation (4) is *in probability* implies that as the tree grows large, the ratio between the occurrences of shapes  $\mathcal{S}^1$  and  $\mathcal{S}^2$  tends to its limit (as opposed to convergence in the mean). This holds for any tree from the process. The variance of the shape occurrence ratio calculated in a group of different trees with the same age  $t$  tends to zero as  $t \rightarrow \infty$ .

This property is very appealing when analyzing large trees, as the difference between the shape occurrence ratio and its limit may be considered statistically insignificant.

In particular, if we are interested in the frequency of  $\mathcal{S}$  (the number of occurrences of shape  $\mathcal{S}$  per tip in the tree), we must divide the number of occurrences of  $\mathcal{S}$  by the number of tips. A tip is also a (very simple) shape  $\mathcal{T}$  defined by the characteristic  $\phi^{\mathcal{T}}(t)$  which equals 1 if  $t$  is larger than the birth time of  $x$ . It follows that  $E[\phi^{\mathcal{T}}(t)] = 1$  for any  $t > 0$ . If the tip characteristic is used in the denominator of equation (4), the denominator is  $1/M$  and we find that the asymptotic frequency of  $\mathcal{S}$  in a tree is

$$M \int_0^\infty e^{-Mt} E[\phi^{\mathcal{S}}(t)] dt, \quad (5)$$

with  $\phi^{\mathcal{S}} := \phi_0^{\mathcal{S}}$  to ensure consistency between the time variables.

## 4 Cherries

### 4.1 Cherries in homogeneous processes

#### Cherry characteristic

At time  $t$ , a cherry  $\mathcal{C}$  is the configuration made by an individual  $x$  and its last descendant  $(x, \xi_x(t))$  when the latter has no descendants, i.e.  $\xi_{(x, \xi_x(t))}(t) = 0$ . It is called a cherry because it visually resembles a cherry (two tips of the tree joined at the same node). To be the ancestor of a cherry,  $x$  must have at least one daughter (i.e.  $\xi_x(t) \geq 1$ ), and  $x$ 's last daughter must have no descendants, i.e.  $\xi_{(x, \xi_x(t))}(t) = 0$ . Accordingly, the cherry characteristic can be written as follows:

$$\phi_x^{\mathcal{C}}(t) = \begin{cases} 1 & \text{if } \xi_x(t) \geq 1 \text{ and } \xi_{(x, \xi_x(t))}(t) = 0 \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

The total number of cherries in the tree is then:  $Z^{\mathcal{C}}(t) = \sum_{x \in I} \phi_x^{\mathcal{C}}(t)$ .

#### Derivation of $E[\phi^{\mathcal{C}}(t)]$

In the homogeneous process, all individuals alive at a given time generate offspring at a constant rate  $\beta$  and die at a constant rate  $\delta$ . In order to derive the asymptotic cherry frequency using equation (5) it is necessary to first derive  $E[\phi^{\mathcal{C}}(t)]$ . For this purpose it is convenient to use the JCCP because in the homogeneous setting the JCCP is the stochastic process that has almost everywhere derivative -1, which jumps at rate  $\beta$  and whose jumps have random size exponentially distributed with parameter  $\delta$  [21]; in addition the process is sent back to the current time  $t$  whenever it overshoots (reflection) and is killed upon hitting 0. Since a tree uniquely defines a JCCP and vice versa, then the law of the JCCP is also the law of the tree (modulo labelling of the tips).

In order to carry out analysis using the JCCP we use  $u_i$  for the  $i$ -th inter-jump time and  $j_{i-1}$  for the  $i$ -th jump size; the notation is depicted in Figures 2 and 3. We aim to find a relationship among the  $u$  and  $j$  variables in the JCCP that is equivalent to the cherry characteristic (6). Recall from section 3 that  $\phi^{\mathcal{C}}(t)$  in Eq. (5) is  $\phi_0^{\mathcal{C}}(t)$ . For this reason we will only focus on the part of the JCCP which corresponds to the ancestor and its last daughter:  $j_0, u_1, j_1, u_2$ .

First note that the ancestor has at least 1 daughter if and only if the JCCP does not hit 0 before the jump  $j_1$ . This ensures the presence of at least one daughter. Secondly, for the last daughter to have no descendants, the second low peak of the JCCP (the base of the third jump  $j_2$ ) has to be lower than the first (the base of the second jump  $j_1$ ).

Recall that the JCCP involves reflection, which means that when the JCCP overshoots, i.e. it jumps beyond the current time threshold  $t$ , the process is sent back to  $t$ . Because reflection may happen on either of the two jumps, there are four distinct cases that give rise to a cherry (see Figure 3). Therefore, a cherry including the ancestor occurs only when the quartet  $(j_0, u_1, j_1, u_2)$  is in one, and

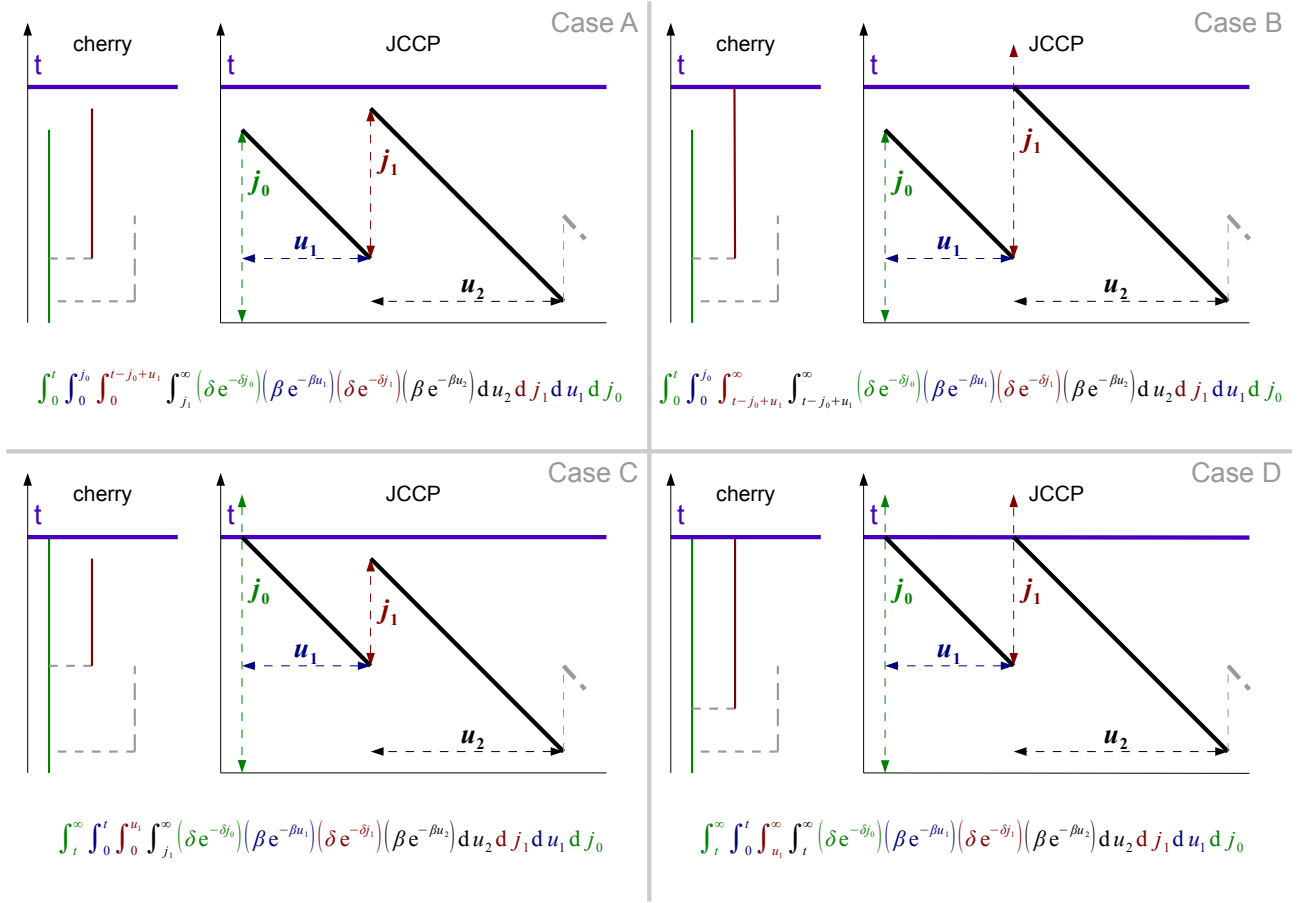


Fig. 3: Schematic for the calculation of  $E[\phi^C(t)]$ . Because each of the two jumps  $j_0$  and  $j_1$  of the JCCP related to a cherry can overshoot above the current time  $t$ , four different cases have to be examined. The probability of each case can be derived with a 4-dimensional integral. In the computation, recall that the  $j_i \sim \text{Exp}(\delta)$  and  $u_i \sim \text{Exp}(\beta)$ , for  $i = 0, 1, 2$ .

one only, of these four subsets of  $\mathbb{R}_+^4$ :

$$A : \begin{cases} 0 < j_0 \leq t \\ 0 < u_1 < j_0 \\ 0 < j_1 \leq t - j_0 + u_1 \\ j_1 < u_2 < \infty \end{cases}, \quad B : \begin{cases} 0 < j_0 \leq t \\ 0 < u_1 < j_0 \\ t - j_0 + u_1 < j_1 < \infty \\ t - j_0 + u_1 < u_2 < \infty \end{cases}, \quad C : \begin{cases} t < j_0 < \infty \\ 0 < u_1 < t \\ 0 < j_1 \leq u_1 \\ j_1 < u_2 < \infty \end{cases}, \quad D : \begin{cases} t < j_0 < \infty \\ 0 < u_1 < t \\ u_1 < j_1 < \infty \\ u_1 < u_2 < \infty \end{cases}.$$

To find the probability that the ancestor mothers a cherry, we are left to measure the four sets above. Recall that  $j_i \sim \text{Exp}(\delta)$  and  $u_i \sim \text{Exp}(\beta)$ , then  $d(j_i) = \delta e^{-\delta j_i} dj_i$  and  $d(u_i) = \beta e^{-\beta u_i} du_i$  for  $i = 0, 1, 2$ . We derive the probability of each set:

$$A : \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{\infty} d(u_2) d(j_1) d(u_1) d(j_0) = \frac{\beta \delta}{(\beta + \delta)^2} - \frac{2\delta}{2\beta + \delta} e^{-\delta t} + \frac{2\delta^2}{(\beta + \delta)^2} e^{-(\beta + \delta)t} - \frac{\beta \delta^2}{(2\beta + \delta)(\beta + \delta)^2} e^{-2(\beta + \delta)t}$$

$$B : \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^{\infty} d(u_2) d(j_1) d(u_1) d(j_0) = \frac{\delta}{2\beta + \delta} e^{-\delta t} - \frac{\delta}{\beta + \delta} e^{-(\beta + \delta)t} + \frac{\beta \delta}{(2\beta + \delta)(\beta + \delta)} e^{-2(\beta + \delta)t}$$

$$C : \int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{\infty} d(u_2) d(j_1) d(u_1) d(j_0) = \frac{\delta}{2\beta + \delta} e^{-\delta t} - \frac{\delta}{\beta + \delta} e^{-(\beta + \delta)t} + \frac{\beta \delta}{(\beta + \delta)(2\beta + \delta)} e^{-2(\beta + \delta)t}$$

$$D : \int_t^\infty \int_0^t \int_{u_1}^\infty \int_{u_1}^\infty d(u_2)d(j_1)d(u_1)d(j_0) = \frac{\beta}{2\beta + \delta} e^{-\delta t} - \frac{\beta}{2\beta + \delta} e^{-2(\beta + \delta)t}$$

Now we only need to sum the four integrals:

$$E[\phi^{\mathcal{C}}(t)] = \frac{\beta\delta}{(\beta + \delta)^2} + \frac{\beta}{2\beta + \delta} e^{-\delta t} - \frac{2\beta\delta}{(\beta + \delta)^2} e^{-(\beta + \delta)t} - \frac{\beta^3}{(2\beta + \delta)(\beta + \delta)^2} e^{-2(\beta + \delta)t} \quad (7)$$

Note that as  $t \rightarrow \infty$ ,  $E[c(t)]$  converges to  $\mathbb{P}(\xi(\infty) \geq 1)\mathbb{P}(\xi(\infty) = 0)$  which is the product of the probabilities of the two events that define the cherry characteristic. Moreover for  $t = 0$  the terms cancel and  $E[c(t)] = 0$ ; this confirms the impossibility to generate a cherry when no time has elapsed.

### The cherries to tips ratio in homogeneous processes

By “cherries to tips ratio” or CTR we indicate the limit of the frequency of the cherry shape  $\mathcal{C}$  in a tree. Using equation (5):

$$\frac{Z^{\mathcal{C}}(t)}{Z^{\mathcal{T}}(t)} \rightarrow CTR = M \int_0^\infty e^{-Mt} E[\phi^{\mathcal{C}}(t)] dt. \quad (8)$$

In homogeneous branching trees the Malthusian parameter is  $M = \beta - \delta$  and it is positive because we assumed the process to be supercritical. The mean number of offspring for each individual, or basic reproduction number, is  $R_0 = \beta/\delta$  which is always greater than 1 under the supercritical assumption. Substituting equation (7) we can compute the integral in equation (8) to obtain  $\frac{\beta}{(3\beta + \delta)(\beta - \delta)}$ . Substitution of  $M$  and  $R_0$  gives

$$CTR = \frac{\beta}{3\beta + \delta} = \frac{R_0}{3R_0 + 1}, \quad (9)$$

Note that in the limit  $\beta \rightarrow \infty$ , that is the limit where the homogeneous tree tends to a Yule tree, the CTR tends to  $1/3$  which is a known result for Yule trees [26, 22]. The methodology used to derive the result in equation (9) can be applied to different shapes and different choices of the branching process (i.e. non-homogeneous). When dealing with non-homogeneous trees we should bear in mind that the JCCP maintains its most important property (independence of  $j_i$  and  $u_i$ ) only when there is a constant birth rate. So for this approach to apply, the non-homogeneity must come from a non-exponential lifespan rather than a non-constant birth rate. Otherwise, evaluation of  $E[\phi^{\mathcal{S}}]$  is likely to be challenging.

## 4.2 Cherries in a non-homogeneous model

Using the same approach as in section 4.1, we found the cherry to tips ratio in a non-homogeneous model. If we choose the life-span distribution to be a Gamma distribution with rate  $\delta$  and shape 2, the Malthusian parameter and  $R_0$  are given by

$$M = \frac{1}{2}\beta - \delta + \frac{1}{2}\sqrt{\beta^2 + 4\beta\delta}, \quad R_0 = \frac{2\beta}{\delta}.$$

Using equation (5) after evaluating  $E[\phi^C(t)]$  in this model, the cherry to tips ratio is:

$$\begin{aligned}
 CTR = & \left( 512 \left( 243R_0^{12} + 243R_0^{23/2} \sqrt{R_0 + 8} + 5103R_0^{11} + 4131R_0^{21/2} \sqrt{R_0 + 8} + 40851R_0^{10} + \right. \right. \\
 & + 26271R_0^{19/2} \sqrt{R_0 + 8} + 160767R_0^9 + 80955R_0^{17/2} \sqrt{R_0 + 8} + 338148R_0^8 + \\
 & + 131184R_0^{15/2} \sqrt{R_0 + 8} + 387448R_0^7 + 112912R_0^{13/2} \sqrt{R_0 + 8} + 235072R_0^6 + \\
 & + 49152R_0^{11/2} \sqrt{R_0 + 8} + 68784R_0^5 + 9360R_0^{9/2} \sqrt{R_0 + 8} + 7616R_0^4 + \\
 & \left. \left. + 512R_0^{7/2} \sqrt{R_0 + 8} + 128R_0^3 \right) \right) \times \\
 & \times \left( 27R_0^5 + 27R_0^{9/2} \sqrt{R_0 + 8} + 297R_0^4 + 189R_0^{7/2} \sqrt{R_0 + 8} + 900R_0^3 + 360R_0^{5/2} \sqrt{R_0 + 8} + \right. \\
 & \left. + 968R_0^2 + 240R_0^{3/2} \sqrt{R_0 + 8} + 368R_0 + 48\sqrt{R_0 + 8}\sqrt{R_0} + 32 \right)^{-1} \times \\
 & \times \left( 3R_0 + \sqrt{R_0 + 8}\sqrt{R_0} \right)^{-2} \times \left( R_0 + \sqrt{R_0 + 8}\sqrt{R_0} \right)^{-2} \times \left( 5R_0 + \sqrt{R_0 + 8}\sqrt{R_0} + 4 \right)^{-3}.
 \end{aligned} \tag{10}$$

## 5 Pitchforks and double-cherries

### Pitchfork characteristics

A pitchfork  $\mathcal{P}$  is a configuration with three tips (illustrated in Figure 4). A pitchfork is formed when an individual  $x'$  last two daughters each have no descendants, or if her last daughter has only one descendant. The characteristic is written accordingly:

$$\phi_x^{\mathcal{P}}(t) = \begin{cases} 1 & \text{if } \begin{cases} \xi_x(t) \geq 2 \text{ and } \xi_{(x, \xi_x(t))}(t) = 0 \text{ and } \xi_{(x, \xi_x(t)-1)}(t) = 0 \\ \text{or} \\ \xi_x(t) \geq 1 \text{ and } \xi_{(x, \xi_x(t))}(t) = 1 \text{ and } \xi_{(x, \xi_x(t), 1)}(t) = 0 \end{cases} \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

### The pitchfork to tips ratio in homogeneous processes

In order to derive the pitchfork to tips ratio (PTR) we use equation (5). To evaluate  $E[\phi^{\mathcal{P}}(t)]$  we use the JCCP as in section 4.1: from the definition in equation (11) we determine the relationships among the first three peaks and three drops of the JCCP that correspond to a pitchfork in the tree. The resulting set in  $\mathbb{R}^6$  is formed of 16 components that are measured with 6-dimensional integrals using the software Maple [1]. There are two cases when a pitchfork is formed: the first extends the cherry and occurs when the ancestor has at least two daughters and the last two of them have no descendants. The second case occurs when the ancestor has at least one daughter and the ancestor's last daughters has one only daughter who does not have any further descendants. Because of possible overshooting, for each case there are 8 different sets that we should measure:

$$\begin{aligned}
 & \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_0^{t-j_0+u_1-j_1+u_2} \int_{j_2}^{\infty} d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
 & \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_{t-j_0+u_1-j_1+u_2}^{\infty} \int_{t-j_0+u_1-j_1+u_2}^{\infty} d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
 & \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_0^{u_2} \int_{j_2}^{\infty} d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
 & \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_{u_2}^{\infty} \int_{u_2}^{\infty} d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
 & \int_0^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_0^{u_1-j_1+u_2} \int_{j_2}^{\infty} d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0)
 \end{aligned}$$



$$\begin{aligned}
& \int_t^\infty \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_{u_1-j_1+u_2}^\infty \int_{u_1-j_1+u_2}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_{u_1}^\infty \int_{u_1}^t \int_0^{u_2} \int_{j_2}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_{u_1}^\infty \int_{u_1}^t \int_{u_2}^\infty \int_{u_2}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_0^{j_1} \int_0^{t-j_0+u_1-j_1+u_2} \int_{t-j_0+u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_0^{j_1} \int_{t-j_0+u_1-j_1+u_2}^\infty \int_{t-j_0+u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^\infty \int_0^{t-j_0+u_1} \int_0^{u_2} \int_{t-u_2+j_2-j_0+u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^\infty \int_0^{t-j_0+u_1} \int_{u_2}^\infty \int_{t-j_0+u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_0^{u_1} \int_0^{j_1} \int_0^{u_1-j_1+u_2} \int_{j_1-u_2+j_2}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_0^{u_1} \int_0^{j_1} \int_{u_1-j_1+u_2}^\infty \int_{u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_{u_1}^\infty \int_0^{u_1} \int_0^{u_2} \int_{u_1-u_2+j_2}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^\infty \int_0^t \int_{u_1}^\infty \int_0^{u_1} \int_{u_2}^\infty \int_{u_1}^\infty d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0)
\end{aligned}$$

To evaluate the integrals, recall that  $j_i \sim \text{Exp}(\delta)$  and  $u_i \sim \text{Exp}(\beta)$ , then  $d(j_i) = \delta e^{\delta j_i} dj_i$  and  $d(u_i) = \beta e^{-\beta u_i} du_i$  for  $i = 0, 1, 2, 3$ . The expression for the pitchfork to tips ratio in the homogeneous model is:

$$\begin{aligned}
PTR &= \frac{3(\beta - \delta)(\beta + \delta)\beta^2}{(2\beta + \delta)(\beta - \delta)(3\beta + \delta)^2} \\
&= \frac{3(R_0 + 1)(R_0^2)}{(2R_0 + 1)(3R_0 + 1)^2}.
\end{aligned} \tag{12}$$

Note that the limit of  $PTR$  as  $R_0 \rightarrow \infty$ , where the homogeneous model tends to the Yule model, is  $1/6$ , which corresponds to Rosenberg's result for the Yule model in [26].

## 5.1 Frequency of the symmetric configuration with 4 tips in homogeneous processes

### The characteristic

Let's consider a configuration  $\mathcal{S}$  of four tips organized in two cherries. In order to count such configurations we define a characteristic which assigns 1 to every individual  $x$  which is both the ancestor and a tip of the configuration. This happens if among the daughters (at least 2) of  $x$ , the last has no descendants and the second last has one only descendant. Equivalently, the ancestor mothers a cherry and so does the ancestor's second last daughter.

$$\phi_x^{\mathcal{S}}(t) = \begin{cases} 1 & \text{if } \xi_x(t) \geq 2 \text{ and } \xi_{(x, \xi_x(t))}(t) = 0 \text{ and } \xi_{(x, \xi_x(t)-1)}(t) = 1 \text{ and } \xi_{(x, \xi_x(t)-1, 1)} = 0 \\ 0 & \text{o.w.} \end{cases}. \tag{13}$$

### Derivation of the asymptotic frequency

We use equation (5) to evaluate the asymptotic frequency of  $\mathcal{S}$ . As in the previous sections we evaluate  $E[\phi^{\mathcal{S}}(t)]$  using the JCCP. Its derivation involves 16 8-dimensional integrals:

$$\begin{aligned}
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_0^{t-j_0+u_1-j_1+u_2} \int_0^{j_2} \int_0^{t-j_0+u_1-j_1+u_2-j_2+u_3} \int_{j_3-u_3+j_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_0^{t-j_0+u_1-j_1+u_2} \int_0^{j_2} \int_{t-j_0+u_1-j_1+u_2-j_2+u_3}^{\infty} \int_{t-j_0+u_1-j_1+u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_{t-j_0+u_1-j_1+u_2}^{\infty} \int_0^{t-j_0+u_1-j_1+u_2} \int_0^{u_3} \int_{t-u_3+j_3-j_0+u_1-j_1+u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_0^{t-j_0+u_1} \int_{j_1}^{j_0-u_1+j_1} \int_{t-j_0+u_1-j_1+u_2}^{\infty} \int_0^{t-j_0+u_1-j_1+u_2} \int_{u_3}^{\infty} \int_{t-j_0+u_1-j_1+u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_0^{u_2} \int_0^{j_2} \int_0^{u_2-j_2+u_3} \int_{j_2-u_3+j_3}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_0^{u_2} \int_0^{j_2} \int_{u_2-j_2+u_3}^{\infty} \int_{u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_{u_2}^{\infty} \int_0^{u_2} \int_0^{u_3} \int_{u_2-u_3+j_3}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_0^t \int_0^{j_0} \int_{t-j_0+u_1}^{\infty} \int_{t-j_0+u_1}^t \int_{u_2}^{\infty} \int_0^{u_2} \int_{u_3}^{\infty} \int_{u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_0^{u_1-j_1+u_2} \int_0^{j_2} \int_0^{u_1-j_1+u_2-j_2+u_3} \int_{j_2-u_3+j_3}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_0^{u_1-j_1+u_2} \int_0^{j_2} \int_{u_1-j_1+u_2-j_2+u_3}^{\infty} \int_{u_1-j_1+u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_{u_1-j_1+u_2}^{\infty} \int_0^{u_1-j_1+u_2} \int_0^{u_3} \int_{u_1-j_1+u_2-u_3+j_3}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_0^{u_1} \int_{j_1}^{t-u_1+j_1} \int_{u_1-j_1+u_2}^{\infty} \int_0^{u_1-j_1+u_2} \int_{u_3}^{\infty} \int_{u_1-j_1+u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_{u_1}^{\infty} \int_{u_1}^t \int_0^{u_2} \int_0^{j_2} \int_0^{u_2-j_2+u_3} \int_{j_2-u_3+j_3}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_{u_1}^{\infty} \int_{u_1}^t \int_0^{u_2} \int_0^{j_2} \int_{u_2-j_2+u_3}^{\infty} \int_{u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_{u_1}^{\infty} \int_{u_1}^t \int_{u_2}^{\infty} \int_0^{u_2} \int_0^{u_3} \int_{t-u_3+j_3-u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0) \\
& \int_t^{\infty} \int_0^t \int_{u_1}^{\infty} \int_{u_1}^t \int_{u_2}^{\infty} \int_0^{u_2} \int_{u_3}^{\infty} \int_{u_2}^{\infty} d(u_4)d(j_3)d(u_3)d(j_2)d(u_2)d(j_1)d(u_1)d(j_0)
\end{aligned}$$

To evaluate the integrals, recall that  $j_i \sim \text{Exp}(\delta)$  and  $u_i \sim \text{Exp}(\beta)$ , then  $d(j_i) = \delta e^{\delta j_i} dj_i$  and  $d(u_i) = \beta e^{-\beta u_i} du_i$  for  $i = 0, 1, 2, 3, 4$ . The asymptotic frequency of  $\mathcal{S}$  is given by:

$$\begin{aligned}
& \frac{1}{4} (2592R_0^9 + 11556R_0^8 + 18279R_0^7 + 13899R_0^6 + 4799R_0^5 - 65R_0^4 - 546R_0^3 - 114R_0^2) \\
& (19440R_0^9 + 91044R_0^8 + 187488R_0^7 + 222741R_0^6 + 168180R_0^5 + 83666R_0^4 + 27416R_0^3 + \\
& + 5705R_0^2 + 684R_0 + 36)^{-1}
\end{aligned} \tag{14}$$

## 6 Conclusion

In this paper we have presented a novel technique to compute the frequency of any shape configuration in a tree. These frequencies, namely the ratio of the number of occurrences of a specific shape configuration to the number of tips in the tree, converge in probability to the expression in Eq. (5) as  $t \rightarrow \infty$ . We have applied the technique to evaluate the asymptotic frequency of cherries in the homogeneous process and in a non-homogeneous process (with non-exponential death/recovery rates). For the homogeneous tree, we have also derived the frequency of pitchforks and of the symmetric shape with four tips (double cherry). In Figure 4 we present a summary of the results.

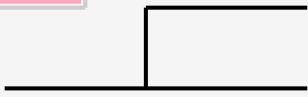
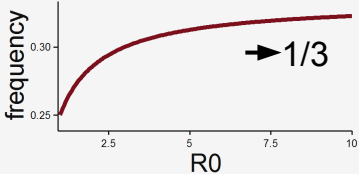

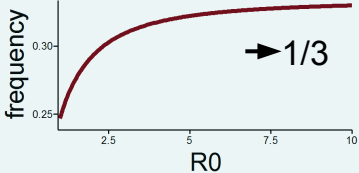
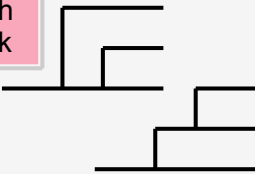
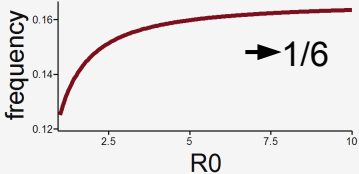
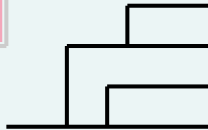
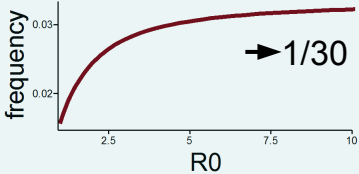
Configuration	Model	As. frequency	Plot and limit
	Homogeneous birth rate: $\beta$ death rate: $\delta$ $R_0 = \beta/\delta$	$\frac{R_0}{3R_0 + 1}$	
	Non homogeneous constant birth rate $\beta$ life span distribution: Gamma (rate= $\delta$ , shape=2) $R_0 = 2\beta/\delta$	$\frac{(512(243R_0^{12} + 243R_0^{11}\sqrt{(R_0+8)} + 5103R_0^{10} + 4131R_0^{9}\sqrt{(R_0+8)} + 40851R_0^8 + 26271R_0^{7/2}\sqrt{(R_0+8)} + 16076R_0^6 + 80955R_0^{5/2}\sqrt{(R_0+8)} + 338148R_0^4 + 131184R_0^{3/2}\sqrt{(R_0+8)} + 387448R_0^3 + 112912R_0^{5/2}\sqrt{(R_0+8)} + 235072R_0^2 + 49152R_0^{3/2}\sqrt{(R_0+8)} + 68784R_0 + 9360R_0^{1/2}\sqrt{(R_0+8)} + 7616R_0 + 512R_0^{3/2}\sqrt{(R_0+8)} + 128R_0))}{((27R_0^4 + 27R_0^3\sqrt{(R_0+8)} + 297R_0^2 + 189R_0\sqrt{(R_0+8)} + 900R_0 + 360R_0^{1/2}\sqrt{(R_0+8)} + 968R_0 + 240R_0^{3/2}\sqrt{(R_0+8)} + 368R_0 + 48\sqrt{(R_0+8)})\sqrt{(R_0+32)} + (3R_0 + \sqrt{(R_0+8)})\sqrt{(R_0+4)})^2 (R_0 + \sqrt{(R_0+8)})\sqrt{(R_0+4)})^{-1}}$	
	Homogeneous birth rate: $\beta$ death rate: $\delta$ $R_0 = \beta/\delta$	$\frac{3R_0^2(R_0 + 1)}{(3R_0 + 1)^2(2R_0 + 1)}$	
	Homogeneous birth rate: $\beta$ death rate: $\delta$ $R_0 = \beta/\delta$	$\frac{1/4(2592R_0^9 + 11556R_0^8 + 18279R_0^7 + 13899R_0^6 + 4799R_0^5 - 65R_0^4 - 546R_0^3 - 114R_0^2 - 19440R_0^9 + 91044R_0^8 + 187488R_0^7 + 222741R_0^6 + 168180R_0^5 + 83666R_0^4 + 27416R_0^3 + 5705R_0^2 + 684R_0 + 36)^{-1}}{(3R_0 + 1)^2(2R_0 + 1)}$	

Fig. 4: Summary of the analytical results.

## References

- [1] Maple 18. Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario 2014.
- [2] ATHREYA, K. B. AND NEY, P. E. (1972). *Branching processes*. Springer.
- [3] BROWN, J. K. (1994). Probabilities of evolutionary trees. *Systematic Biology* **43**, 78–91.
- [4] CAVALLI-SFORZA, L. L. AND EDWARDS, A. W. (1967). Phylogenetic analysis. models and estimation procedures. *American journal of human genetics* **19**, 233.
- [5] CHANG, H. AND FUCHS, M. (2010). Limit theorems for patterns in phylogenetic trees. *Journal of Mathematical Biology* **60**, 481–512.
- [6] COLIJN, C. AND GARDY, J. (2014). Phylogenetic tree shapes resolve disease transmission patterns. *Evolution, Medicine, and Public Health* **2014**, 96–108.
- [7] COX, D. R. (1962). Renewal theory.
- [8] DIDELOT, X., GARDY, J. AND COLIJN, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution* **31**, 1869–1879.
- [9] DRUMMOND, A. J. AND RAMBAUT, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.
- [10] EDWARDS, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society. Series B (Methodological)* 155–174.

- [11] FROST, S. D. AND VOLZ, E. M. (2013). Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**,.
- [12] GEIGER, J. (1995). Contour processes of random trees. *London mathematical society lecture note series* 72–96.
- [13] GERNHARD, T., HARTMANN, K. AND STEEL, M. (2008). Stochastic properties of generalised yule models, with biodiversity applications. *Journal of Mathematical Biology* **57**, 713–735.
- [14] HARDING, E. (1971). The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability* 44–77.
- [15] HOLMES, E. C., NEE, S., RAMBAUT, A., GARNETT, G. P. AND HARVEY, P. H. (1995). Revealing the history of infectious disease epidemics through phylogenetic trees. *Philosophical Transactions of the Royal Society B: Biological Sciences* **349**, 33–40.
- [16] JAGERS, P. (1969). Renewal theory and the almost sure convergence of branching processes. *Arkiv för Matematik* **7**, 495–504.
- [17] JAGERS, P. (1975). *Branching Processes with Biological Applications*. Wiley.
- [18] KATO-MAEDA, M., HO, C., PASSARELLI, B., BANAEI, N., GRINSDALE, J., FLORES, L., ANDERSON, J., MURRAY, M., ROSE, G., KAWAMURA, L. M. ET AL. (2013). Use of whole genome sequencing to determine the microevolution of mycobacterium tuberculosis during an outbreak. *PLOS ONE* **8**, e58235.
- [19] LAMBERT, A. (2008). Population dynamics and random genealogies. *Stochastic Models* **24**, 45–163.
- [20] LAMBERT, A., ALEXANDER, H. K. AND STADLER, T. (2014). Phylogenetic analysis accounting for age-dependent death and sampling with applications to epidemics. *Journal of Theoretical Biology* **352**, 60–70.
- [21] LAMBERT, A. ET AL. (2010). The contour of splitting trees is a lévy process. *The Annals of Probability* **38**, 348–395.
- [22] MCKENZIE, A. AND STEEL, M. (2000). Distributions of cherries for two models of trees. *Mathematical Biosciences* **164**, 81–92.
- [23] NERMAN, O. (1981). On the convergence of supercritical general (cmj) branching processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57**, 365–395.
- [24] PAGE, R. D. (1991). Random dendrograms and null hypotheses in cladistic biogeography. *Systematic Biology* **40**, 54–62.
- [25] POON, A. F., WALKER, L. W., MURRAY, H., MCCLOSKEY, R. M., HARRIGAN, P. R. AND LIANG, R. H. (2013). Mapping the shapes of phylogenetic trees from human and zoonotic rna viruses. *PLOS ONE* **8**, e78122.
- [26] ROSENBERG, N. A. (2006). The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. *Annals of Combinatorics* **10**, 129–146.
- [27] STADLER, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology* **261**, 58–66.
- [28] WILSON, D. J., FALUSH, D. AND MCVEAN, G. (2005). Germs, genomes and genealogies. *Trends in Ecology & Evolution* **20**, 39–45.
- [29] YPMA, R. J., VAN BALLEGOIJEN, W. M. AND WALLINGA, J. (2013). Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062.